TREC 2013 Web Track Overview

Kevyn Collins-Thompson
University of Michigan

Paul Bennett, Fernando Diaz
Microsoft Research

Charles L. A. Clarke
University of Waterloo

Ellen M. Voorhees
NIST

January 30, 2014

1 Introduction

The goal of the TREC Web track is to explore and evaluate retrieval approaches over large-scale subsets of the Web – currently on the order of one billion pages. For TREC 2013, the fifth year of the Web track, we implemented the following significant updates compared to 2012.

First, the Diversity task was replaced with a new *Risk-sensitive retrieval* task that explores the tradeoffs systems can achieve between effectiveness (overall gains across queries) and robustness (minimizing the probability of significant failure, relative to a provided baseline). Second, we based the 2013 Web track experiments on the new ClueWeb12 collection created by the Language Technologies Institute at Carnegie Mellon University. ClueWeb12 is a successor to the ClueWeb09 dataset, comprising about one billion Web pages crawled between Feb-May 2012.[1] The crawling and collection process for ClueWeb12 included a rich set of seed URLs based on commercial search traffic, Twitter and other sources, and multiple measures for flagging undesirable content such as spam, pornography, and malware. The Adhoc task continued as in previous years.

Both the Adhoc and Risk-sensitive tasks used a common topic set of 50 new topics, and differed only in their evaluation methodology. With the goal of reflecting aspects of authentic Web usage, the Web track topics were again developed from the logs and data resources of commercial search engines. However, a different extraction methodology was used compared to last year. This year, two types of topics were developed: faceted topics,

[1] Details on ClueWeb12 are available at `http://boston.lti.cs.cmu.edu/clueweb12`

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
30 JAN 2014		**00-00-2014 to 00-00-2014**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
TREC 2013 Web Track Overview	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Michigan,Ann Arbor,MI,48109**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
proceedings of the Twenty-Second Text REtrieval Conference (TREC 2013) held in Gaithersburg, Maryland, November 19-22, 2013. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) and the Defense Advanced Research Projects Agency (DARPA).

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**15**	

Task	Adhoc	Risk	Total
Groups	4	11	15
Runs	34	27	61

Table 1: TREC 2013 Web Track participation.

and unfaceted (single-facet) topics. Faceted topics were more like "head" queries, and structured as having a representative set of subtopics, with each subtopic corresponding to a popular subintent of the main topic. The faceted topic queries were less directly ambiguous than last year: the ambiguity lay more in which subintents were likely to be most relevant to users and not in the direct interpretation of the query. Unfaceted (single-facet) topics were intended to be more like "tail" queries with a clear question or intent. For faceted topics, query clusters were developed and used by NIST for topic development. Only the base query was released to participants initially: the topic structures containing subtopics and single- vs multi-faceted *vs.* topic type were only released after runs were submitted. This was done to avoid biases that might be caused by revealing extra information about the information need that may not be available to Web search systems as part of the actual retrieval process.

The Adhoc task judged documents with respect to the topic as a whole. Relevance levels are similar to the levels used in commercial Web search, including a spam/junk level. The top two levels of the assessment structure are related to the older Web track tasks of homepage finding and topic distillation. Subtopic assessment was also performed for the faceted topics, as described further in Section 3.

Table 1 summarizes participation in the TREC 2013 Web Track. Overall, we received 61 runs from 15 groups: 34 ad hoc runs and 27 risk-sensitive runs. The number of participants in the Web track increased slightly over 2012 (when 12 groups participated, submitting 48 runs), including some institutions that had not previously participated. Eight runs were manual runs, submitted across four groups: all other runs were automatic with no human intervention. Nine of the runs used the Category B subset of ClueWeb12: all other runs used the main Category A corpus.

2 ClueWeb12 Category A and B collections

As with ClueWeb09, the ClueWeb12 collection comes with two datasets: Category A, and Category B. The Category A dataset is the main corpus and contains about 733 million documents (27.3 TB uncompressed, 5.54 TB compressed). The Category B dataset is a sample from Category A, containing about 52 million documents, or about 7% of the Category A total. Details on how the Category A and B collections were created may be found on the Lemur project website[2]. We strongly encouraged participants to use the full Category A data set if possible. Results in the paper are labeled by their collection category.

3 Topics

NIST created and assessed 50 new topics for the Web track. Unlike TREC 2012, the TREC 2013 Web track included a significant proportion of more focused topics designed to represent more specific, less frequent, possibly more difficult queries. To retain the Web flavor of queries in this track, we retain the notion from last year that some topics may be multi-faceted, i.e. broader in intent and thus structured as a representative set of subtopics, each related to a different potential aspect of user need. Examples are provided below. For topics with multiple subtopics, documents were judged with respect to the subtopics. For each subtopic, NIST assessors made a scaled six-point judgment as to whether or not the document satisfied the information need associated with the subtopic. For those topics with multiple subtopics, the set of subtopics was intended to be representative, not exhaustive.

Subtopics were based on information extracted from the logs of a commercial search engine. Topics having multiple subtopics had subtopics selected roughly by overall popularity, which was achieved using combined query suggestion and completion data from two commercial search engines. In this way, the focus was retained on a balanced set of popular subtopics, while limiting the occurrence of strange and unusual interpretations of subtopic aspects. Single-facet topic candidates were developed based on queries extracted from search log data that were low-frequency ('tail-like') but issued by multiple users; less than 10 terms in length; and relatively low effectiveness scores across multiple commercial search engines (as of January 2013).

The topic structure was similar to that used for the TREC 2009 topics.

[2] `http://lemurproject.org/clueweb12/specs.php`

Examples of single-facet topics include:

```
<topic number="227" type="single">
<query>i will survive lyrics</query>
<description>Find the lyrics to the song "I Will Survive".</description>
</topic>

<topic number="229" type="single">
<query>beef stroganoff recipe</query>
<description>
Find complete (not partial) recipes for beef stroganoff.
</description>
</topic>
```

Examples of faceted topics include:

```
<topic number="235" type="faceted">
<query>ham radio</query>
<description>How do you get a ham radio license?</description>
<subtopic number="1" type="inf">How do you get a ham radio license?</subtopic>
<subtopic number="2" type="nav">What are the ham radio license classes?</subtopic>
<subtopic number="3" type="inf">How do you build a ham radio station?</subtopic>
<subtopic number="4" type="inf">Find information on ham radio antennas.</subtopic>
<subtopic number="5" type="nav">What are the ham radio call signs?</subtopic>
<subtopic number="6" type="nav">Find the web site of Ham Radio Outlet.</subtopic>
</topic>

<topic number="245" type="faceted">
<query>roosevelt island</query>
<description>What restaurants are on Roosevelt Island (NY)?</description>
<subtopic number="1" type="inf">What restaurants are on Roosevelt Island (NY)?</subtopic>
<subtopic number="2" type="nav">Find the Roosevelt Island tram schedule.</subtopic>
<subtopic number="3" type="inf">What is the history of the Roosevelt Island tram?</subtopic>
<subtopic number="4" type="nav">Find a map of Roosevelt Island (NY).</subtopic>
<subtopic number="5" type="inf">
Find real estate listings for Roosevelt Island (NY).
</subtopic>
```

Initial topic release to participants included only the query field, as shown in the excerpt here:

```
201:raspberry pi
202:uss carl vinson
203:reviews of les miserables
```

```
204:rules of golf
205:average charitable donation
```

As shown in the above examples, those topics with a clear focused intent have a single subtopic. Topics with multiple subtopics reflect underspecified queries, with different aspects covered by the subtopics. We assume that a user interested in one aspect may still be interested in others. Each subtopic was categorized as being either navigational ("nav") or informational ("inf"). A navigational subtopic usually has only a small number of relevant pages (often one). For these subtopics, we assume the user is seeking a page with a specific URL, such as an organization's homepage. On the other hand, an informational query may have a large number of relevant pages. For these subtopics, we assume the user is seeking information without regard to its source, provided that the source is reliable.

For the adhoc task, relevance is judged on the basis of the description field. Thus, the first subtopic is always identical to the description sentence.

4 Methodology and Measures

4.1 Pooling and Judging

For each topic, participants in the adhoc and risk-sensitive tasks submitted a ranking of the top 10,000 results for that topic. All submitted runs were included in the pool for judging. A common pool was used.

For the risk-sensitive task, new versions of `ndeval` and `gdeval` that supported the risk-sensitive versions of the evaluation measures (described below) were provided to NIST.

All data and tools required for evaluation, including the scoring programs ndeval and gdeval as well as the baseline run used in computation of the risk-sensitive scores (run results-cata-filtered.txt) were available in the track's github distribution[3].

The relevance judgment for a page was one of a range of values as described in Section 4.2. The topic-aspect combinations with zero known relevant documents were eliminated from all the evaluations. These are:

topic 202, aspect 2
topic 202, aspect 3
topic 216, aspect 2
topic 225, aspect 1
topic 225, aspect 5

[3] `http://github.com/trec-web/trec-web-2013`

topic 244, aspect 2
topic 244, aspect 3

For topics that had a single aspect in the original topics file, that one aspect is used. For all other topics except topic 225, aspect number 1 is the single aspect. For topic 225, aspect number 2 is the single aspect (aspect 2 is used instead of aspect 1 because aspect 1 has no known relevant documents).

Different topics were pooled to different depths because the original depth (20) resulted in too many documents to be judged in the allotted amount of assessing time. Smaller pools were built using a depth of 10. In all cases, the pools were built over all submitted runs. Pools were sorted so that pages from the same site (as determined by URL syntax) were contiguous in the pool. For multi-aspect topics, assessors judged a given page against all aspects before moving to the next page in the pool. Topics judged to depth 20 were: 201, 202, 203, 204, 205, 206, 208, 210, 211, 212, 214, 215, 216, 217, 218, 219, 220, 221, 223, 224, 225, 226, 232, 233, 234, 239, 240, 243, 247. Topics judged to depth 10 were: 207, 209, 213, 222, 227, 228, 229, 230, 231, 235, 236, 237, 238, 241, 242, 244, 245, 246, 248, 249, 250.

4.2 Ad-hoc Retrieval Task

An ad-hoc task in TREC provides the basis for evaluating systems that search a static set of documents using previously-unseen topics. The goal of an ad-hoc task is to return a ranking of the documents in the collection in order of decreasing probability of relevance. The probability of relevance for a document is considered independently of other documents that appear before it in the result list. For the ad-hoc task, documents are judged on the basis of the description field using a six-point scale, defined as follows:

1. Nav: This page represents a home page of an entity directly named by the query; the user may be searching for this specific page or site. (relevance grade 4)

2. Key: This page or site is dedicated to the topic; authoritative and comprehensive, it is worthy of being a top result in a web search engine. (relevance grade 3)

3. HRel: The content of this page provides substantial information on the topic. (relevance grade 3)

4. Rel: The content of this page provides some information on the topic, which may be minimal; the relevant information must be on that page,

not just promising-looking anchor text pointing to a possibly useful page. (relevance grade 1)

5. Non: The content of this page does not provide useful information on the topic, but may provide useful information on other topics, including other interpretations of the same query. (relevance grade 0)

6. Junk: This page does not appear to be useful for any reasonable purpose; it may be spam or junk (relevance grade -2).

After each description we list the relevance grade assigned to that level as they appear in the judgment (qrels) file. These relevance grades are also used for calculating graded effectiveness measures, except that a value of -2 is treated as 0 for this purpose. For binary effectiveness measures, we treat grades 1/2/3/4 as relevant and grades 0/-2 as non-relevant.

The primary effectiveness measure for the ad-hoc task is *expected reciprocal rank* (ERR) as defined by Chapelle et al. [1]. We also report a variant of NDCG [3] as well as standard binary effectiveness measures, including mean average precision (MAP) and precision at rank k (P@k). To account for the faceted topics, we also report diversity-based versions of these measures: intent-aware expected reciprocal rank (ERR-IA) [1] and α-nDCDG [2].

Figure 1 summarizes the per-topic variability in ERR@10 across all submitted runs. Figure 2 shows the variability in ERR@10 for two specific top-ranked runs, from the Technion and University of Glasgow, with baseline included for comparison.

4.3 Risk-sensitive Retrieval Task

The new *risk-sensitive retrieval* task for Web evaluation rewards algorithms that not only achieve improvements in average effectiveness across topics (as in the ad-hoc task), but also maintain good robustness, which we define as *minimizing the risk of significant failure* relative to a given baseline.

Search engines use increasingly sophisticated stages of retrieval in their quest to improve result quality: from personalized and contextual re-ranking to automatic query reformulation. These algorithms aim to increase retrieval effectiveness on average across queries, compared to a baseline ranking that does not use such operations. However, these operations are also risky since they carry the possibility of failure – that is, making the results worse than if they had not been used at all. The goal of the risk-sensitive task is two-fold: 1) To encourage research on algorithms that go beyond just optimizing average effectiveness in order to effectively optimize both effectiveness and ro-

bustness, and achieve effective tradeoffs between these two competing goals; and 2) to explore effective risk-aware evaluation criteria for such systems.

The risk-sensitive retrieval track is related to the goals of the earlier TREC Robust Track (TREC 2004, 2005),[4] which focused on increasing retrieval effectiveness for poorly-performing topics using evaluation measures such as geometric MAP that focused on maximizing the average improvement on the most difficult topics. The risk-sensitive retrieval track can be thought of as a next step in exploring more general retrieval objectives and evaluation measures that (a) explicitly account for, and can differentiate systems based on, differences in *variance* or other risk-related statistics of the win/loss distribution across topics for a single run, (b) the quality of *the curve derived from a set of tradeoffs* between effectiveness and robustness achievable by systems, measured across *multiple runs* at different average effectiveness levels, and (c) computing (a) and (b) by accounting for the effectiveness of a competing baseline (both standard, and participant-supplied) as a factor in optimizing retrieval performance.

As a standard baseline, we used a pseudo-relevance feedback run as implemented by the Indri retrieval engine.[5] Specifically, for each query, we used 10 feedback documents, 20 feedback terms, and a linear interpolation weight of 0.60 with the original query. Additionally, we used the Waterloo spam classifier to filter out all documents with a percentile-score less than 70.[6].

As with the adhoc task, we use Intent-Aware Expected Reciprocal Rank (ERR-IA) as the basic measure of retrieval effectiveness, and per-query retrieval delta is defined as the absolute difference in effectiveness between a contributed run and the above standard baseline run, for a given query. A positive delta means a win for the system on that query, and negative delta means a loss. We also report other flavors of the risk-related measure based on NDCG. For single runs, the following will be the main risk-sensitive evaluation measure. Let $\Delta(q) = R_A(q) - R_{BASE}(q)$ be the absolute win or loss for query q with system retrieval effectiveness $R_A(q)$ relative to the baseline's effectiveness $R_{BASE}(q)$ for the same query. We categorize the outcome for each query q in the set Q of all N queries according to the sign of $\Delta(q)$, giving three categories:
Hurt Queries (Q_-) have $\Delta(q) < 0$; Unchanged Queries (Q_0) have $\Delta(q) = 0$; Improved Queries (Q_+) have $\Delta(q) > 0$.

[4] `http://trec.nist.gov/data/robust/04.guidelines.html`
[5] `http://www.lemurproject.org/indri/`
[6] `http://www.mansci.uwaterloo.ca/~msmucker/cw12spam/`

The risk-sensitive utility measure $U_{RISK}(Q)$ of a system over the set of queries Q is defined as:

$$U_{RISK}(Q) = 1/N \cdot [\Sigma_{q \in Q_+} \Delta(q) - (\alpha + 1)\Sigma_{q \in Q_-} \Delta(q)] \quad (1)$$

where α is the key risk-aversion parameter. In words, this rewards systems that maximize average effectiveness, but also penalizes losses relative to the baseline results for the same query, weighting losses $\alpha + 1$ times as heavily as successes. When the risk aversion parameter α is large, a system will become more conservative and put more emphasis on avoiding large losses relative to the baseline. When α is small, a system will tend to ignore the baseline. The adhoc task objective, maximizing only average effectiveness across queries, corresponds to the special case $\alpha = 0$. Details are given in Appendix A of the TREC Web 2013 Guidelines[7].

Table 2: Top ad-hoc task results ordered by ERR@10. Only the best run according to ERR@10 from each group is included in the ranking.

Group	Run	Cat	Type	ERR@10	nDCG@10
Technion	clustmrfaf	A	auto	0.175	0.298
udel_fang	UDInfolabWEB2	A	auto	0.167	0.284
uogTr	uogTrAIwLmb	A	auto	0.151	0.247
udel	udelManExp	A	manual	0.150	0.241
ICTNET	ICTNET13RSR2	A	auto	0.149	0.224
ut	ut22xact	A	auto	0.144	0.230
diro_web_13	udemQlm1lFbWiki	A	auto	0.143	0.255
wistud	wistud.runD	A	manual	0.125	0.215
CWI	cwiwt13cps	A	auto	0.121	0.211
UJS	UJS13LCRAd2	B	auto	0.100	0.155
webis	webisrandom	A	auto	0.093	0.171
RMIT	RMITSC75	A	auto	0.093	0.172
Organizers	baseline	A	auto	0.088	0.162
MSR_Redmond	msr_alpha0_95_4	A	manual	0.087	0.157
UWaterlooCLAC	UWCWEB13RISK02	A	auto	0.080	0.134
DLDE	dlde	B	manual	0.008	0.009

[7] http://research.microsoft.com/en-us/projects/trec-web-2013/

Table 3: Top ad-hoc task results ordered by ERR@20. Only the best run according to ERR@20 from each group is included in the ranking.

Group	Run	Cat	Type	ERR@20	nDCG@20
Technion	clustmrfaf	A	auto	0.184	0.310
udel_fang	UDInfolabWEB2	A	auto	0.176	0.282
uogTr	uogTrAIwLmb	A	auto	0.160	0.259
ICTNET	ICTNET13RSR2	A	auto	0.158	0.236
udel	udelManExp	A	manual	0.157	0.246
ut	ut22xact	A	auto	0.152	0.228
diro_web_13	udemQlm1lFbWiki	A	auto	0.152	0.254
wistud	wistud.runD	A	manual	0.134	0.225
CWI	cwiwt13cps	A	auto	0.128	0.218
UJS	UJS13LCRAd2	B	auto	0.107	0.148
RMIT	RMITSCTh	A	auto	0.102	0.179
webis	webisrandom	A	auto	0.101	0.181
MSR_Redmond	msr_alpha0_95_4	A	manual	0.097	0.175
Organizers	baseline	A	auto	0.096	0.168
UWaterlooCLAC	UWCWEB13RISK02	A	auto	0.085	0.132
DLDE	dlde	B	manual	0.008	0.007

5 Conclusions and Future Plans

The Web track will continue for a sixth year in TREC 2014, using substantially the same tasks and methodology as this year, but with potential adjustments in some aspects. The following are known areas for refinement, based on participant feedback and our experience organizing this year's Web track.

- Improved methodology for developing and assessing the more focused, unfaceted (also referred to as single-facet) topics. Many of the unfaceted topic candidates were indeed unambiguous, more tail-like queries, but a number had potentially multiple answers (e.g. [dark chocolate health benefits]). This led to many pages being partially relevant, with no clear way for assessors to know when it was complete enough. We believe retaining a blend of more or less focused query types is important to reflect the nature of authentic Web queries, but will look at revised query clusters and clearer topic development and assessment guidelines for unfaceted topics.

Table 4: Top diversity measures on results ordered by ERR-IA@10. Only the best run according to ERR-IA@10 from each group is included in the ranking.

Group	Run	Cat	Type	ERR-IA@10	α-nDCG@10	NRBP
udel_fang	UDInfolabWEB2	A	auto	0.574	0.628	0.547
Technion	clustmrfaf	A	auto	0.554	0.620	0.521
ICTNET	ICTNET13RSR3	A	auto	0.542	0.598	0.512
uogTr	uogTrAIwLmb	A	auto	0.539	0.606	0.498
udel	udelPseudo2	A	auto	0.531	0.612	0.486
ut	ut22base	A	auto	0.506	0.574	0.470
wistud	wistud.runD	A	manual	0.503	0.558	0.466
diro_web_13	udemQlm1lFbWiki	A	auto	0.475	0.557	0.433
CWI	cwiwt13cps	A	auto	0.473	0.531	0.439
UJS	UJS13Risk2	B	auto	0.461	0.516	0.434
webis	webismixed	A	auto	0.409	0.468	0.374
RMIT	RMITSC75	A	auto	0.376	0.448	0.330
MSR_Redmond	msr_alpha0_95_4	A	manual	0.358	0.444	0.306
Organizers	baseline	A	auto	0.342	0.416	0.294
UWaterlooCLAC	UWCWEB13RISK02	A	auto	0.315	0.373	0.283
DLDE	dlde	B	manual	0.045	0.058	0.038

- Having a two-stage submission process to determine baselines, or user-supplied baselines. This would involve ad-hoc runs being submitted first, followed by selection of some of those runs to be redistributed to participants to use in the 2nd re-ranking task.
- Using judgments more effectively/judging more queries - one idea that might be feasible if we do a two-stage process is to specifically target queries where the submitted ad-hoc runs have high variability. It would require careful thought since it is possible the risk-sensitive approaches could still degrade performance here.
- More holistic types of analysis that compare tradeoff curves within and across systems.

We will continue the use of ClueWeb12 as the test collection for TREC 2014.

Table 5: Top diversity measures ordered by ERR-IA@20. Only the best run according to ERR-IA@20 from each group is included in the ranking.

Group	Run	Cat	Type	ERR-IA@20	α-nDCG@20	NRBP
udel_fang	UDInfolabWEB2	A	auto	0.582	0.654	0.547
Technion	clustmrfaf	A	auto	0.567	0.668	0.521
ICTNET	ICTNET13RSR3	A	auto	0.551	0.627	0.512
uogTr	uogTrAIwLmb	A	auto	0.548	0.637	0.498
udel	udelPseudo2	A	auto	0.539	0.637	0.486
ut	ut22base	A	auto	0.513	0.596	0.470
wistud	wistud.runD	A	manual	0.512	0.589	0.466
CWI	cwiwt13cps	A	auto	0.480	0.557	0.439
diro_web_13	udemQlm1lFbWiki	A	auto	0.480	0.576	0.433
UJS	UJS13Risk2	B	auto	0.468	0.539	0.434
webis	webismixed	A	auto	0.423	0.516	0.374
RMIT	RMITSCTh	A	auto	0.388	0.489	0.330
MSR_Redmond	msr_alpha1	A	manual	0.368	0.476	0.308
Organizers	baseline	A	auto	0.352	0.451	0.294
UWaterlooCLAC	UWCWEB13RISK02	A	auto	0.323	0.399	0.283
DLDE	dlde	B	manual	0.045	0.058	0.038

6 Acknowledgements

We thank Jamie Callan, David Pane and the Language Technologies Institute at Carnegie Mellon University for creating and distributing the ClueWeb12 collection. This track could not operate without this valuable resource. Also, thanks to Nick Craswell for his many valuable suggestions and feedback.

References

[1] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 621–630, New York, NY, USA, 2009. ACM.

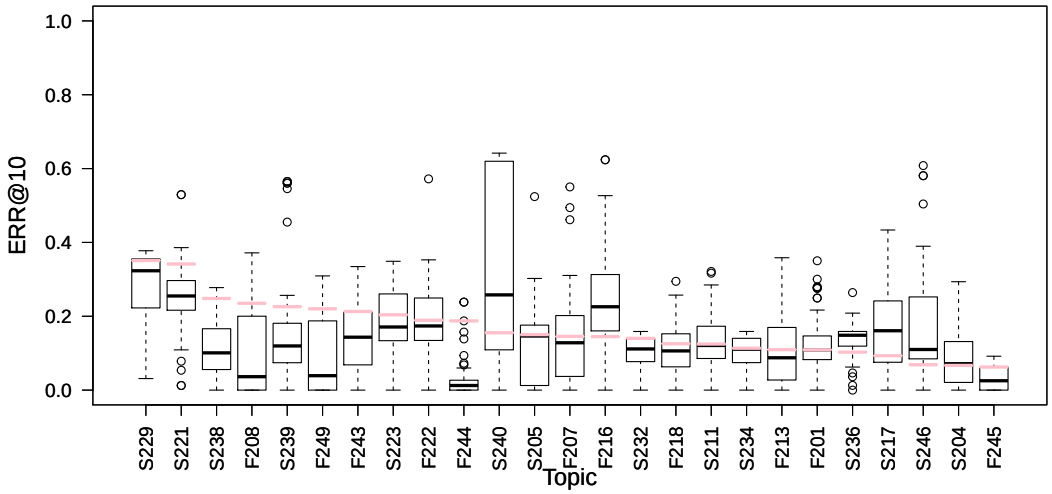
[2] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM*

Table 6: Overall ERR@10 and risk measures for each team according to difference from the baseline's ERR@10 ("Organizers" below). Ordered by $\alpha = 1$ performance (*i.e.*, slight risk sensitivity). The best performance in each column was selected for a team and therefore this may be overly optimistic.

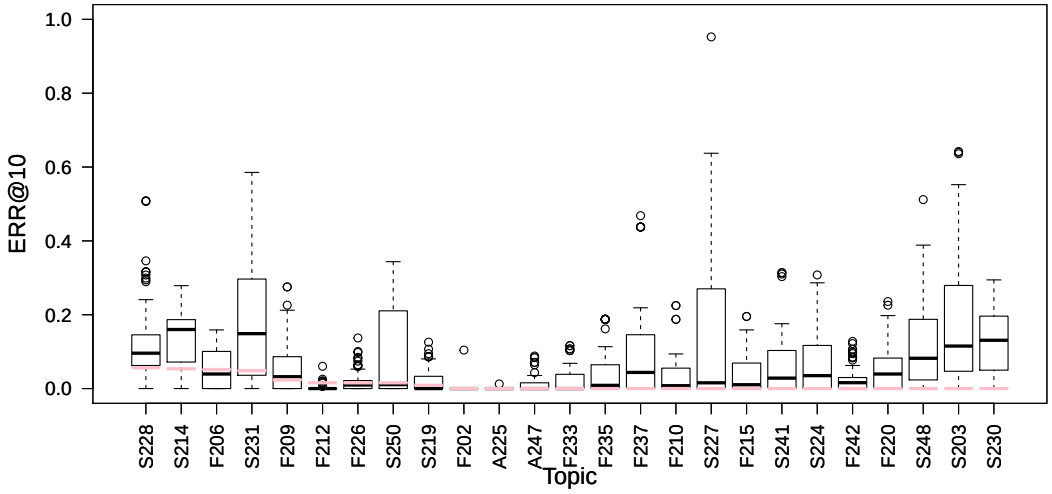
Group	ERR@10	$\Delta, \alpha = 0$	$\Delta, \alpha = 1$	$\Delta, \alpha = 5$	$\Delta, \alpha = 10$
Technion	0.175	0.087	0.076	0.033	-0.020
udel_fang	0.167	0.078	0.059	-0.018	-0.114
udel	0.150	0.061	0.047	-0.011	-0.084
diro_web_13	0.143	0.055	0.034	-0.051	-0.158
uogTr	0.151	0.062	0.030	-0.101	-0.265
ICTNET	0.149	0.060	0.028	-0.079	-0.209
ut	0.144	0.056	0.025	-0.098	-0.248
wistud	0.125	0.037	0.005	-0.063	-0.143
CWI	0.121	0.033	0.003	-0.115	-0.263
Organizers	0.088	0.000	0.000	0.000	0.000
MSR_Redmond	0.087	-0.001	-0.009	-0.042	-0.084
RMIT	0.093	0.005	-0.027	-0.156	-0.317
UJS	0.100	0.012	-0.027	-0.184	-0.379
webis	0.093	0.005	-0.029	-0.163	-0.332
UWaterlooCLAC	0.080	-0.009	-0.040	-0.164	-0.319
DLDE	0.008	-0.081	-0.162	-0.486	-0.891

SIGIR conference on Research and development in information retrieval, SIGIR '08, pages 659–666, New York, NY, USA, 2008. ACM.

[3] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, Oct. 2002.

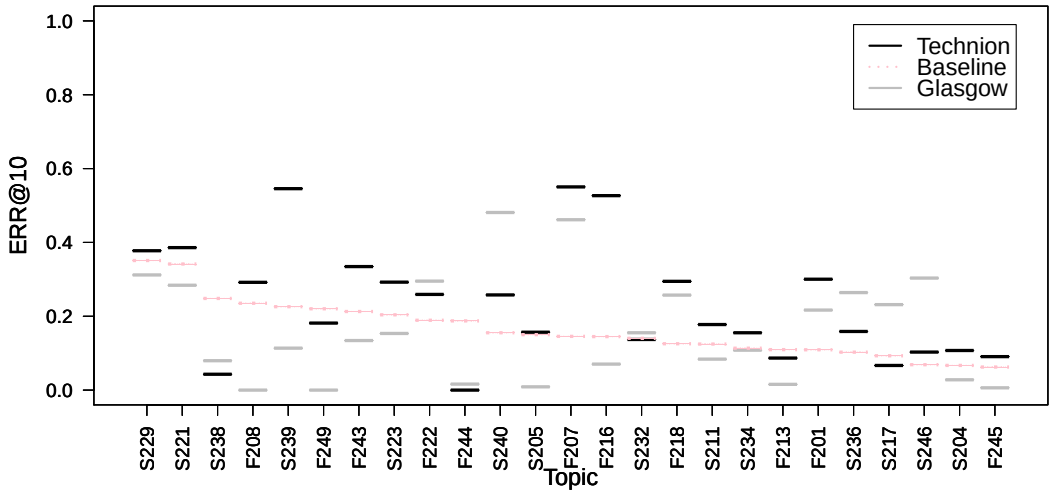


(a) Top 25 topics (by descending baseline ERR@10)

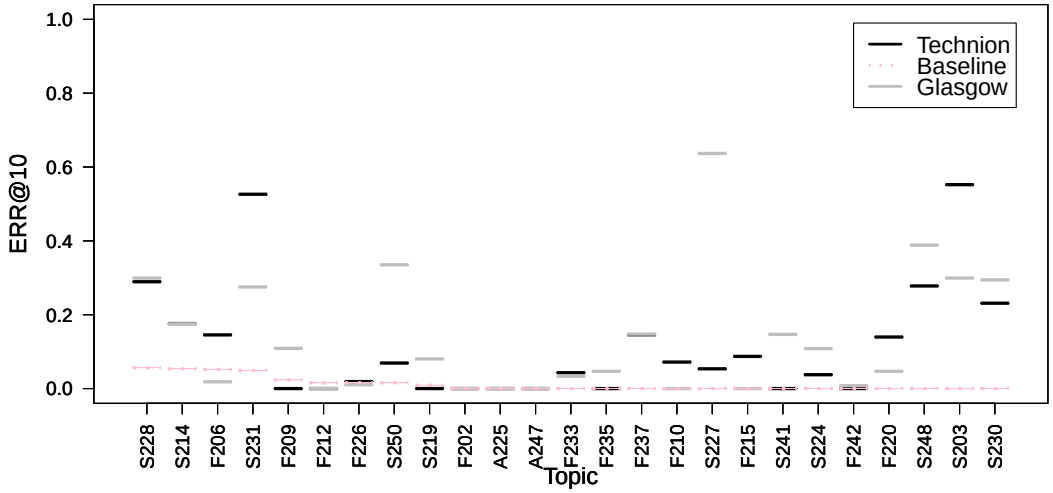


(b) Bottom 25 topics (by descending baseline ERR@10

Figure 1: Boxplots for TREC 2013 Web topics, showing variation in ERR@10 effectiveness across all submitted runs. Topics are sorted by decreasing baseline ERR@10 (pink bar). Faceted topics are prefixed with 'F', single-facet topics by 'S', ambigous topics by 'A' .



(a) Top 25 topics (by descending baseline ERR@10)



(b) Bottom 25 topics (by descending baseline ERR@10)

Figure 2: Chart showing the significant variation in ERR@10 per topic between top-ranked Technion and Glasgow runs. Topics are sorted by decreasing baseline ERR@10 (pink dashed bar). Faceted topics are prefixed with 'F', single-facet topics by 'S', ambigous topics by 'A'.